A survey of corpora for Germanic low-resource languages and dialects

Verena Blaschke, Hinrich Schütze & Barbara Plank
LMU Munich

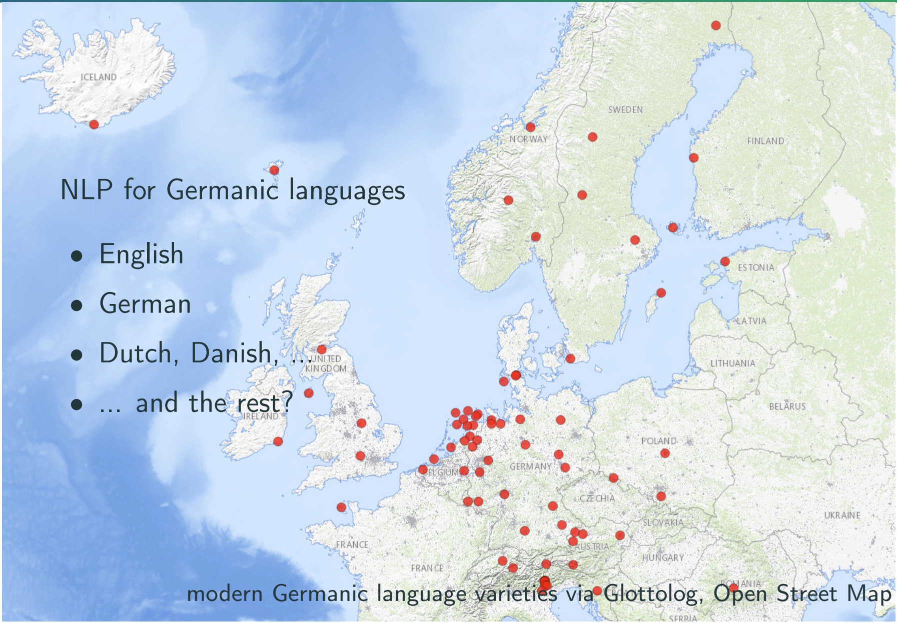
May 23, 2023 @ NoDaLiDa,
Tórshavn (Faroe Islands)

Motivation

NLP for Germanic languages

- English
- German
- Dutch, Danish, ...

Motivation



NLP for Germanic languages

- English
- German
- Dutch, Danish, ...
- ... and the rest?

modern Germanic language varieties via Glottolog, Open Street Map

Motivation

- Recent surveys of other LRL resources, like creoles (Lent et al., 2022), regional languages of Italy (Ramponi, 2022), Ethiopia (Tonja et al., 2023), ...
- General focus on the challenges that come with NLP for LRLs

Motivation

Why do we care?

- Access to language technology
- Linguistics research
- Research on ML in low-resource, high-variability scenarios
- Many of the corpora in this overview were released in the context of specific NLP research

...but it's hard to get started with any of these if you can't find the datasets that already exist!

This talk

- How to find corpora? Which types were we interested in?
- Language varieties
- How do you represent data from unwritten/newly written languages?
- Annotation types
- Curation types & data quality
- Recommendations

How do you find corpora?

- Publications (ACL Anthology, arXiV)
  → a lot manual digging...
- Dataset repos (Zenodo, CLARIN, LRE Map, ...)

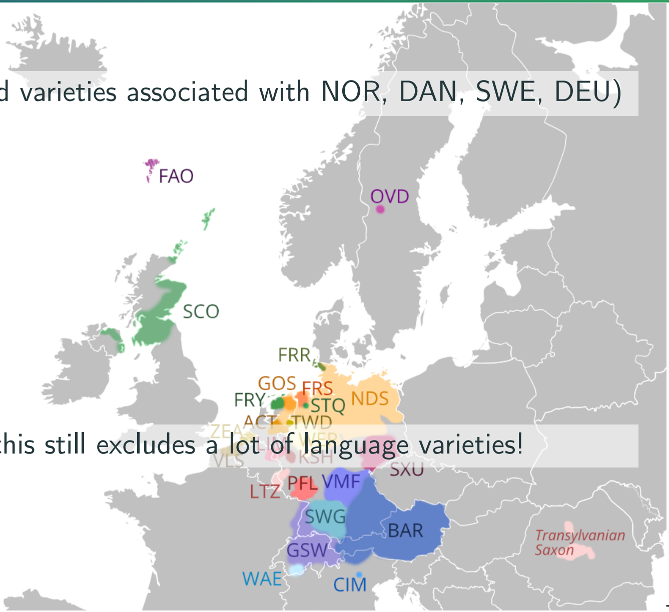
Summer/fall 2022 and early 2023

Inclusion criteria

- Recent (rather than historic) data
- No creoles/pidgins
- Full sentences/utterances (no word lists)
- (In)directly accessible for research (*and* the accessibility status is indicated!)
- Computer-friendly formats (XML, TSV, TXT, ...) rather than PDF etc.
- High-quality data (look out for, e.g., OCR issues!)

→80+ corpora!

For which language varieties did we find datasets?



(+non-standard varieties associated with NOR, DAN, SWE, DEU)

...but this still excludes a lot of language varieties!

How do you represent a (mostly spoken) language?

A challenge/consideration relevant to many corpora!

- Privacy issues concerning audio material
- Wildly different transcription/orthography styles

How do you represent a (mostly spoken) language?

¶ Normalized text (closely related standard language)

✎ Phone[m/t]ic transcriptions

NB Tale

¶ Etter litt godsnakk kom tre av kyrne mot han
mens den fjerde glei og fall

✎ ""{t@4 l"it g""u:snAkk k"Om t4"e: "A:v C"y:n‘@ m"u:t "An
m"ens d_= fj""{:d‘@ gl"eI "O: f"Al

✎ ^{2}ɛtəɹ ^{1}lɪt ^{2}guːsnɑkk ^{1}kɔm ^{1}tɹeː ^{1}ɑːv ^{1}çyːɳə ^{1}muːt ^{1}ɑn
^{1}mɛns dn̩ ^{2}fjæːɖə ^{1}glɛɪ̯ ^{1}oː ^{1}fɑl

“After some coaxing, three of the cows came towards him
while the fourth slipped and fell.”

How do you represent a (mostly spoken) language?

✎ Transcription styles can vary a lot!

ArchiMob

¶ können sie ihre jugendzeit beschreiben

✎ chönd sii iri jugendziit beschriibe

“Can you describe your youth?”

How do you represent a (mostly spoken) language?

A Some LRLs have (more or less! accepted) standardized orthographies

Nordic Dialect Corpus

A [...] wen war eð før ien månað ? juni ?

¶ [...] vad var det för en månad ? juni ?

“[...] what month was it? June?”

UD Low Saxon LSDC

A Nu leyt em de böyse vynd disse nacht kyn ouge an enander doon.

✎ Nu leit em de baise Find düse Nacht kinn Auge an enander dohn.

“Now the wicked enemy didn’t let them get a wink of sleep that night.”

✎ ... and speakers might just use idiosyncratic ad-hoc spellings!

Annotations

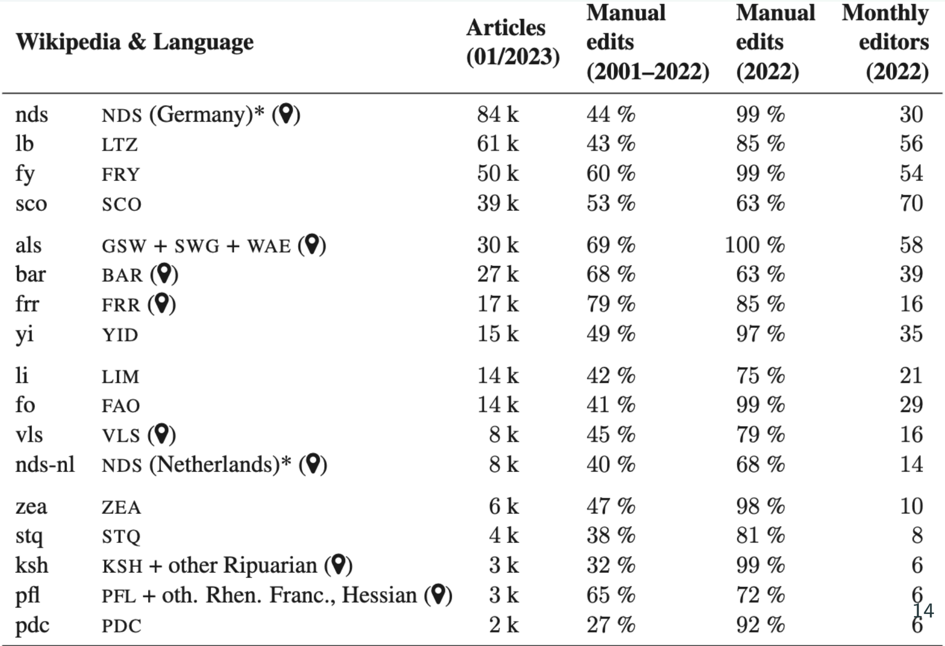
What, if any, high-quality annotations do we find?

- Many: not annotated
- $\sim 1/3$: Geolocation, dialect group (especially common for audio corpora)
- $\sim 1/5$: Morphosyntax (POS tags, dependencies, phrase structure; 6 UD corpora, varying tag sets)
- Rare: 3 × (manual, curated) translations, 1 × paraphrases, 1 × sentiment, 1 × topics, 1 × slot and intent detection

(Un)curated data & data quality

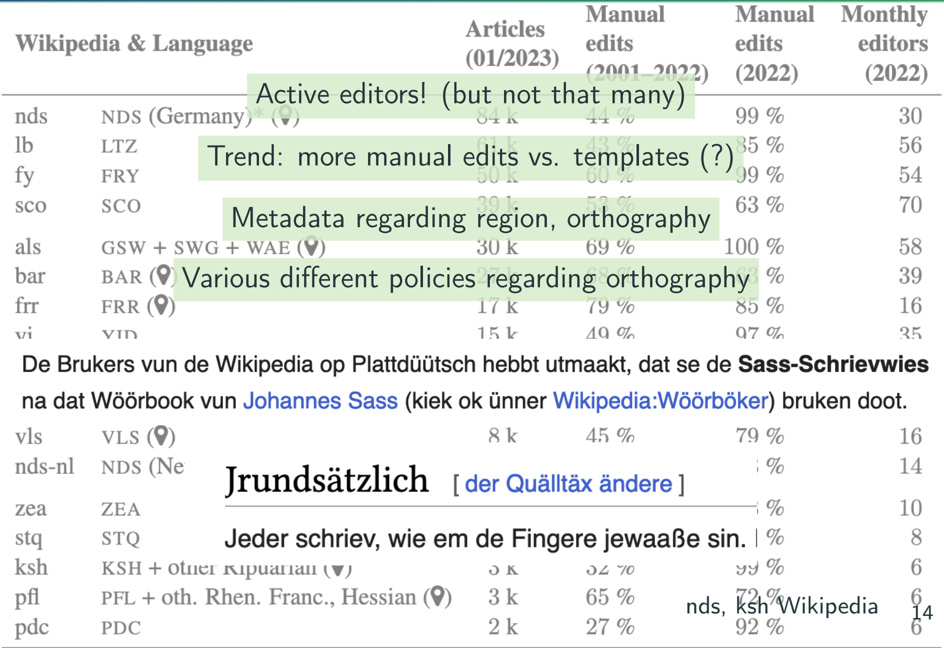
- Mostly ($> 80\,\%$) curated datasets (elicited, transcribed, from books, manually checked web data, ...)
- Also some based on webcrawls or community-based data collection efforts
  - Some dedicated webcrawls (SwissSpot, ...)
  - CommonCrawl and versions thereof (OSCAR, CC-100)
  - Parallel data: Tatoeba, open-source software
  - Wikipedia

(Un)curated data & data quality



Wikipedia & Language		Articles (01/2023)	Manual edits (2001–2022)	Manual edits (2022)	Monthly editors (2022)
nds	NDS (Germany)* (📍)	84 k	44 %	99 %	30
lb	LTZ	61 k	43 %	85 %	56
fy	FRY	50 k	60 %	99 %	54
sco	SCO	39 k	53 %	63 %	70
als	GSW + SWG + WAE (📍)	30 k	69 %	100 %	58
bar	BAR (📍)	27 k	68 %	63 %	39
frr	FRR (📍)	17 k	79 %	85 %	16
yi	YID	15 k	49 %	97 %	35
li	LIM	14 k	42 %	75 %	21
fo	FAO	14 k	41 %	99 %	29
vls	VLS (📍)	8 k	45 %	79 %	16
nds-nl	NDS (Netherlands)* (📍)	8 k	40 %	68 %	14
zea	ZEA	6 k	47 %	98 %	10
stq	STQ	4 k	38 %	81 %	8
ksh	KSH + other Ripuarian (📍)	3 k	32 %	99 %	6
pfl	PFL + oth. Rhen. Franc., Hessian (📍)	3 k	65 %	72 %	6
pdc	PDC	2 k	27 %	92 %	6

(Un)curated data & data quality



Wikipedia & Language		Articles (01/2023)	Manual edits (2001–2022)	Manual edits (2022)	Monthly editors (2022)
nds	NDS (Germany)* (📍)	84 k	44 %	99 %	30
lb	LTZ	61 k	43 %	85 %	56
fy	FRY	50 k	60 %	99 %	54
sco	SCO	39 k	53 %	63 %	70
als	GSW + SWG + WAE (📍)	30 k	69 %	100 %	58
bar	BAR (📍)	27 k	68 %	63 %	39
frr	FRR (📍)	17 k	79 %	85 %	16
yi	YID	15 k	49 %	97 %	35
vls	VLS (📍)	8 k	45 %	79 %	16
nds-nl	NDS (Ne[illegible]	[illegible]	[illegible]	[illegible] %	14
zea	ZEA	[illegible]	[illegible]	[illegible] %	10
stq	STQ	[illegible]	[illegible]	[illegible] %	8
ksh	KSH + other Ripuarian (📍)	5 k	32 %	99 %	6
pfl	PFL + oth. Rhen. Franc., Hessian (📍)	3 k	65 %	72 %	6
pdc	PDC	2 k	27 %	92 %	6

Active editors! (but not that many)

Trend: more manual edits vs. templates (?)

Metadata regarding region, orthography

Various different policies regarding orthography

De Brukers vun de Wikipedia op Plattdüütsch hebbt utmaakt, dat se de **Sass-Schrievwies** na dat Wöörbook vun Johannes Sass (kiek ok ünner Wikipedia:Wöörböker) bruken doot.

Jrundsätzlich [der Quälltäx ändere]

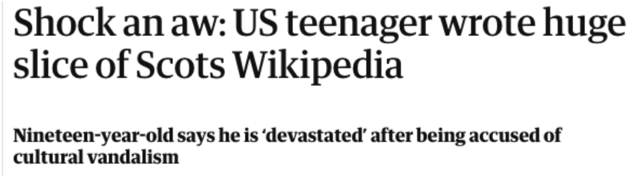
Jeder schriev, wie em de Fingere jewaaße sin.

nds, ksh Wikipedia

(Un)curated data & data quality

Look at the raw data from uncurated sources!

- Uncurated LRL data tend to be rather low quality (wrong language, bad data cleaning) (Kreutzer et al., 2022; Abadji et al., 2022)
- Low-status varieties prone to parodies?



Shock an aw: US teenager wrote huge slice of Scots Wikipedia

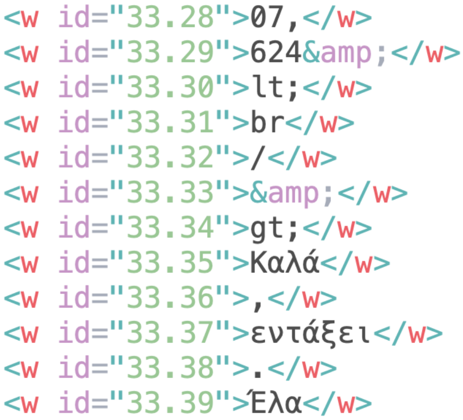
Nineteen-year-old says he is 'devastated' after being accused of cultural vandalism

Brooks & Hern, 2020, The Guardian

(Un)curated data & data quality

You don't always need to speak the language to identify quality issues

"West Flemish" *QED OPUS*
(automatically constructed from crowd-sourcing data)



```
<w id="33.28">07,</w>
<w id="33.29">624&amp;</w>
<w id="33.30">lt;</w>
<w id="33.31">br</w>
<w id="33.32">/</w>
<w id="33.33">&amp;</w>
<w id="33.34">gt;</w>
<w id="33.35">Καλά</w>
<w id="33.36">,</w>
<w id="33.37">εντάξει</w>
<w id="33.38">.</w>
<w id="33.39">Έλα</w>
```

Summary

- Not a lot of datasets *per language*, but 80+ in total
- Largely unannotated
- Mostly dedicated, curated corpora

Recommendations

… for using LRL corpora

- Check the quality!
- Check whether your (pre-)training, dev, and test data are truly from independent sources (datasets overlap!)
- Also look for quantitative works by dialectologists and sociolinguists (outside traditional NLP venues)

Recommendations

… for creating LRL corpora

- Document the transcription guidelines / orthographies (if applicable)
- Use archives geared towards long-term storage (CLARIN, LRE Map, Zenodo)
- Share basic metadata like corpus size, data sources, annotation procedure + license information

Future work

… for anyone who's interested :)

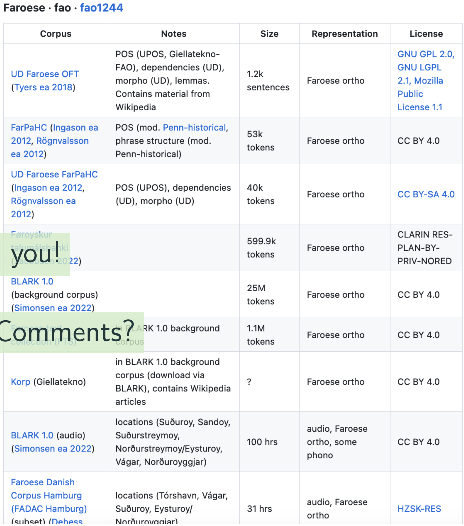
- Non-corpus resources like word lists – which ones are out there and how can we integrate them into our research?
- Language models & other resources

github.com/mainlp/
germanic-lrl-corpora

(To be continually updated!)

Thank you!

Questions? Comments?



Faroese · fao · fao1244

Corpus	Notes	Size	Representation	License
UD Faroese OFT (Tyers ea 2018)	POS (UPOS, Giellatekno-FAO), dependencies (UD), morpho (UD), lemmas. Contains material from Wikipedia	1.2k sentences	Faroese ortho	GNU GPL 2.0, GNU LGPL 2.1, Mozilla Public License 1.1
FarPaHC (Ingason ea 2012, Rögnvalsson ea 2012)	POS (mod. Penn-historical, phrase structure (mod. Penn-historical)	53k tokens	Faroese ortho	CC BY 4.0
UD Faroese FarPaHC (Ingason ea 2012, Rögnvalsson ea 2012)	POS (UPOS), dependencies (UD), morpho (UD)	40k tokens	Faroese ortho	CC BY-SA 4.0
Føroyskur talumálsbanki (Jacobsen 2022)		599.9k tokens	Faroese ortho	CLARIN RES-PLAN-BY-PRIV-NORED
BLARK 1.0 (background corpus) (Simonsen ea 2022)		25M tokens	Faroese ortho	CC BY 4.0
Faroese text collection (FTS)	in BLARK 1.0 background corpus	1.1M tokens	Faroese ortho	CC BY 4.0
Korp (Giellatekno)	in BLARK 1.0 background corpus (download via BLARK), contains Wikipedia articles	?	Faroese ortho	CC BY 4.0
BLARK 1.0 (audio) (Simonsen ea 2022)	locations (Suðuroy, Sandoy, Suðurstreymoy, Norðurstreymoy/Eysturoy, Vágar, Norðuroyggjar)	100 hrs	audio, Faroese ortho, some phono	CC BY 4.0
Faroese Danish Corpus Hamburg (FADAC Hamburg) (subset) (Debess	locations (Tórshavn, Vágar, Suðuroy, Eysturoy/ Norðuroyggjar)	31 hrs	audio, Faroese ortho	HZSK-RES

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot (2022). "Towards a cleaner document-oriented multilingual crawled corpus." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4344–4355. European Language Resources Association.

Julia Kreutzer, Isaac Caswell, et al. (2022). "Quality at a glance: An audit of web-crawled multilingual datasets." *Transactions of the Association for Computational Linguistics*, 10:50–72.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard (2022). "What a creole wants, what a creole needs." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6439–6449. European Language Resources Association.

Alan Ramponi (2022). "NLP for language varieties of Italy: Challenges and the path forward." *Computing Research Repository*, arXiv:2209.09757.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam (2023). "Natural language processing in Ethiopian languages: Current state, challenges, and opportunities." In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL)*.